



# INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

## AI-Based Automatic Text Recognition And Text-To-Sound Conversion

<sup>1</sup>Kalathma M K, <sup>2</sup>S N Varshith, <sup>3</sup>Shashank B C, <sup>4</sup>Shivaprasad H M, <sup>5</sup>Tejas S

<sup>1</sup>Assistant Professor, <sup>2</sup>Student, <sup>3</sup>Student, <sup>4</sup>Student, <sup>5</sup>Student

<sup>1</sup>Department of Computer science and Engineering,

<sup>1</sup>ATME college of Engineering, Mysuru, India

**Abstract:** With the rapid advancement of artificial intelligence, there is a growing demand for assistive technologies that enhance accessibility. This paper presents an AI-Based Automatic Text Recognition and Text-to-Sound Conversion System designed to extract text from images and convert it into natural speech. The system integrates Google Cloud Vision API for Optical Character Recognition (OCR) and Google Cloud Text-to-Speech API for speech synthesis, enabling real-time text processing and audio generation. A key feature of this system is its multilingual support, allowing automatic detection and processing of text in multiple languages without manual selection. Users can upload images or capture them via a webcam, and the extracted text is converted into speech, which can be played back or downloaded in MP3 format. The backend is developed using Django, ensuring scalability, security, and ease of integration. This project aims to enhance digital accessibility by providing a seamless solution for users with visual impairments or reading difficulties. By leveraging AI-driven OCR and TTS technologies, it simplifies text digitization and promotes inclusivity. The proposed system demonstrates real-time efficiency, high accuracy, and adaptability, making it a practical tool for education, content accessibility, and assistive applications.

**Key words:** Optical Character Recognition (OCR), Text-to-Speech (TTS), Google Cloud Vision API, Accessibility, Django, AI-Based Speech Conversion.

### I. INTRODUCTION

Artificial intelligence has significantly advanced, leading to an increasing need for assistive technologies that enhance accessibility. The AI-Based Automatic Text Recognition and Text-to-Sound Conversion System seamlessly extracts text from images and converts it into speech, providing a practical solution for visually impaired users, language learners, and individuals requiring digital content conversion. The system integrates Google Cloud Vision API for Optical Character Recognition (OCR) and Google Cloud Text-to-Speech API for high-quality speech synthesis. Users can upload images or capture text in real-time via a webcam, and the system processes the text, detects the language, and converts it into natural-sounding speech in MP3 format. Developed using the Django web framework, the system ensures robust backend functionalities, secure user authentication, and scalable deployment. Additional features include multilingual support, real-time text extraction, and an intuitive user interface for seamless user interaction.

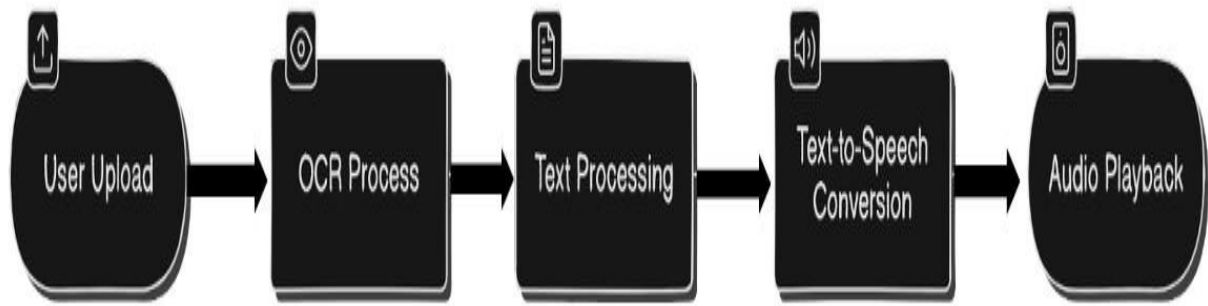


Fig: Functional Flow of Text-to-Speech Application.

Conventional Text-to-Speech (TTS) applications require users to input text manually, making them inefficient when dealing with printed documents or handwritten text. Many OCR tools operate separately from TTS systems, leading to fragmented workflows and requiring third-party software for text extraction. Moreover, existing OCR systems often struggle with handwritten text recognition and multilingual support, limiting their effectiveness in diverse linguistic environments. These limitations necessitate an integrated, automated solution that efficiently extracts and vocalizes text from various sources. Despite significant advancements in OCR and TTS technologies, major challenges persist in real-time processing, multilingual support, and seamless integration with speech synthesis tools. Users frequently face delays, fragmented workflows, and manual language selection requirements, affecting accessibility and ease of use. This project aims to bridge this gap by developing an AI-powered system capable of real-time text recognition from images, automatic language detection to support diverse scripts, high-quality speech conversion without manual intervention, and user-friendly interaction through a web-based interface. The proposed system integrates OCR and TTS functionalities into a single, automated workflow, eliminating the need for multiple tools. Users can upload images or capture text using their webcam, and the system extracts the text, detects its language, and converts it into speech output in MP3 format. Key features of the system include image upload and capture supporting static image uploads and live camera captures via a web-based interface, text detection and extraction utilizing Google Cloud Vision API to extract text from both printed and handwritten documents, multilingual support for automatic detection and processing of English, Hindi, and Kannada, real-time text-to-speech conversion generating natural, high-quality speech output for immediate playback or download, user authentication providing secure login, registration, and password recovery features using Django's authentication system, error handling and user feedback detecting errors such as unreadable text or image processing failures and providing real-time feedback, and scalability and robustness built using Django, allowing future enhancements, API integration, and large-scale deployment. The proposed system offers several advantages over conventional OCR and TTS applications. Unlike existing systems requiring separate OCR and TTS tools, this system combines both in a single platform. It ensures real-time text extraction and speech conversion, eliminating delays. Language support is significantly improved, with automatic detection of English, Hindi, and Kannada. The system effectively recognizes both printed and handwritten text, unlike traditional OCR solutions that struggle with handwriting. Additionally, it provides secure user authentication, including OTP-based password recovery, and is built on Django for scalability and maintainability, supporting future enhancements. By addressing the shortcomings of manual text input, slow processing, and fragmented workflows, this project demonstrates a practical, AI-driven approach to improving text accessibility. The combination of Django, Google Cloud APIs, and real-time processing makes this system a valuable tool for various applications, including education, accessibility, and digital content conversion.

## II. METHODOLOGY

The project integrates **Optical Character Recognition (OCR)** and **Text-to-Speech (TTS)** technologies to process images containing text and convert them into speech. It utilizes **Google Cloud Vision API** for text recognition and **Google Cloud Text-to-Speech API** for speech generation.

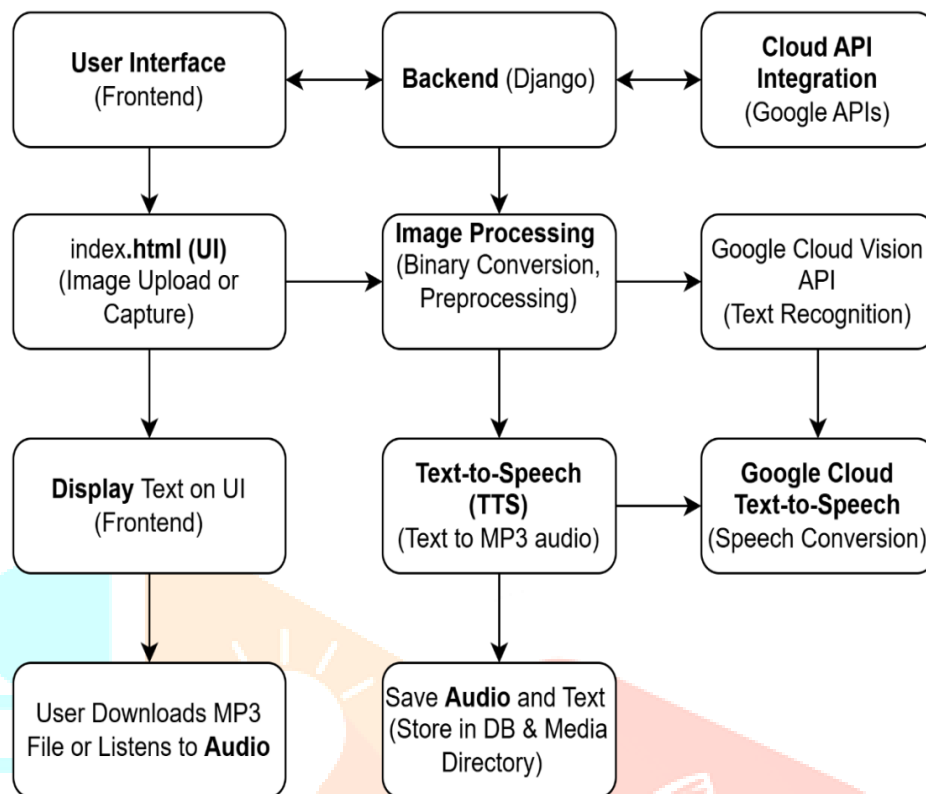


Fig: System Methodology Block Diagram showing the flow from user input to text recognition using Google Cloud Vision API, followed by text-to-speech conversion using Google Cloud Text-to-Speech API.

### Step 1: User Registration, Login, and Authentication

- User Registration (Sign Up)
  - Users enter their username, email, and password.
  - The system validates unique email addresses to prevent duplication.
  - Passwords are encrypted and securely stored using Django's authentication methods.
- User Login (Sign In)
  - Users input their email and password.
  - The system verifies credentials using Django authentication.
  - Upon successful login, access is granted; otherwise, an error message is displayed.
- Password Reset
  - Users can request a password reset if they forget their credentials.
  - An OTP is sent to their registered email for verification.
  - Upon OTP validation, the user can securely set a new password.

## Step 2: Image Upload and Processing

- **Uploading Images**
  - Users can upload images via an **HTML form**.
  - The image is **converted into binary data** for backend processing.
- **Capturing Images via Webcam**
  - Users can **capture images using a webcam**.
  - The image is **encoded in base64** format and then converted to binary data.
- **Image Preprocessing**
  - Before text recognition, the image undergoes the following preprocessing steps:
    - **Grayscale Conversion** – Converts color images to grayscale for better contrast.
    - **Thresholding (Binarization)** – Converts images to black and white for improved text detection.
    - **Noise Reduction** – Removes irrelevant artifacts to enhance clarity.
    - **Resizing** – Adjusts the image dimensions to meet **OCR input requirements**.

## Step 3: Text Recognition Using Google Cloud Vision API

- **Text Detection**
  - The **Google Cloud Vision API** detects text regions in the image.
  - The system locates, extracts, and analyses the text.
- **Text Recognition Using Convolutional Neural Networks (CNNs)**
  - The API applies **CNN models** to recognize text patterns.
  - The model processes the image through multiple layers to handle variations in **font styles, handwriting, and distortions**.
  - The final **text output** is extracted for further processing.

## Step 4: Algorithm Text-to-Speech Conversion Using Google Cloud Text-to-Speech API

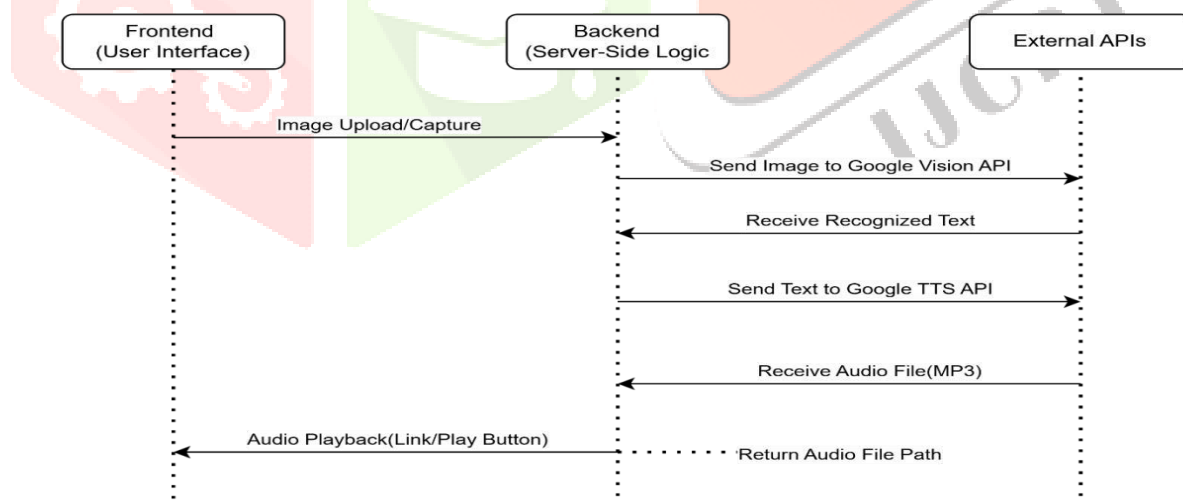
- **Sending Text to TTS API**
  - The extracted text is **sent to the Google Cloud TTS API**.
  - The API's **SynthesizeSpeech()** method is used for speech generation.
- **Speech Synthesis Using WaveNet Neural Network**
  - The API uses **WaveNet**, a deep neural network, to create **realistic speech**.
  - The text is **converted into phonemes** (small sound units).
- **Generating and Playing the Audio**
  - The phonemes are **synthesized into an audio waveform**.
  - The output is saved as an **MP3 file** and provided to the user for playback.

## Step 5: Integration of Frontend, Backend, and APIs

- **Frontend-Backend Communication**
  - Users upload/capture images on the **frontend**.
  - The frontend **sends image data** to the backend.
  - The backend processes the image, extracts text and converts it into **speech**.
  - The final **audio file is sent back** to the frontend for playback.
- **Backend-API Interaction**
  - The backend communicates with **Google Cloud Vision API** to recognize text.
  - The extracted text is sent to **Google Cloud Text-to-Speech API** for speech conversion.

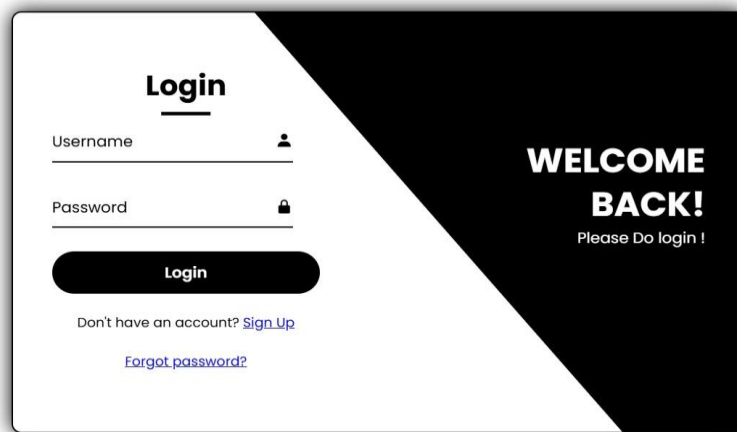
## Step 6: Challenges and Solutions

- **Handling Image Quality Variations**
  - Low-quality images, lighting issues, and distortions affected OCR accuracy.
  - **Solution:** Applied **image preprocessing techniques** like resizing, noise reduction, and binarization to enhance text clarity.
- **Supporting Multiple Languages**
  - Some users required text recognition and speech synthesis in **different languages**.
  - **Solution:** Implemented **language detection**, ensuring the appropriate language model is used for **speech synthesis**.
- **Optimizing API Usage and Costs**
  - Excessive API calls increased **costs and usage limitations**.
  - **Solution:** Optimized API calls by **reducing redundant requests** and implementing **result caching** for efficiency.



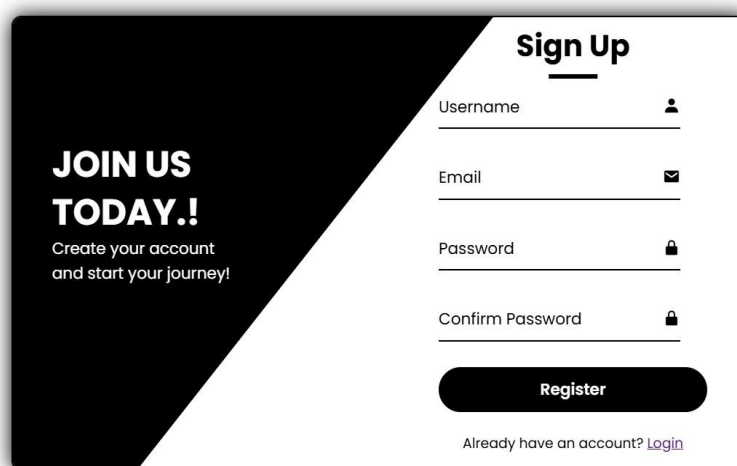
**Fig:** Block Diagram for Frontend, Backend, and API Integration.

### III. RESULT



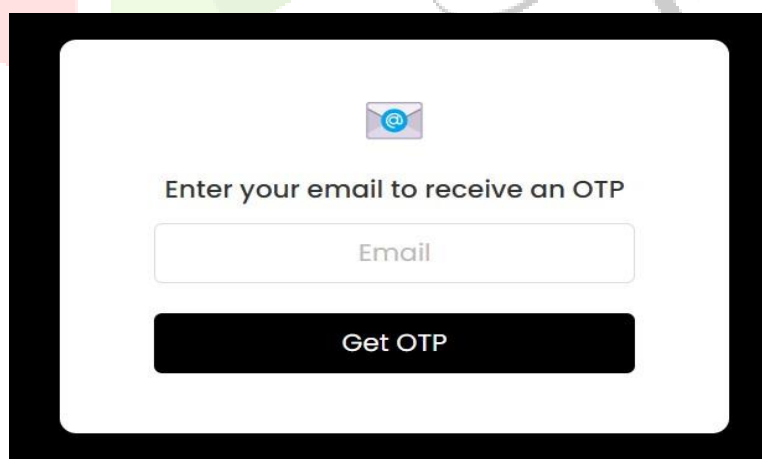
The login page features a dark blue header with the title "Login" in white. Below the title, there are two input fields: "Username" with a person icon and "Password" with a lock icon. A black "Login" button is positioned below the password field. At the bottom, there are two links: "Don't have an account? [Sign Up](#)" and "[Forgot password?](#)". On the right side, a white box contains the text "WELCOME BACK!" in bold, followed by "Please Do login !".

Fig: The login page allows users to authenticate using their registered credentials.



The sign up page has a dark blue header with the title "Sign Up" in white. On the left, a white box contains the text "JOIN US TODAY.!" in bold, followed by "Create your account and start your journey!". On the right, there are four input fields: "Username" with a person icon, "Email" with an envelope icon, "Password" with a lock icon, and "Confirm Password" with a lock icon. A black "Register" button is located below the "Confirm Password" field. At the bottom, there is a link: "Already have an account? [Login](#)".

Fig: The signup page enables new users to register by providing their details. Duplicate username or email inputs trigger error messages.



The password reset page has a dark blue header. Below the header, there is a white box with a blue envelope icon at the top. The text "Enter your email to receive an OTP" is centered. Below the text is a white input field with the placeholder text "Email". At the bottom of the white box is a black "Get OTP" button.

Fig: Password reset page where users enter their email to get an OTP.

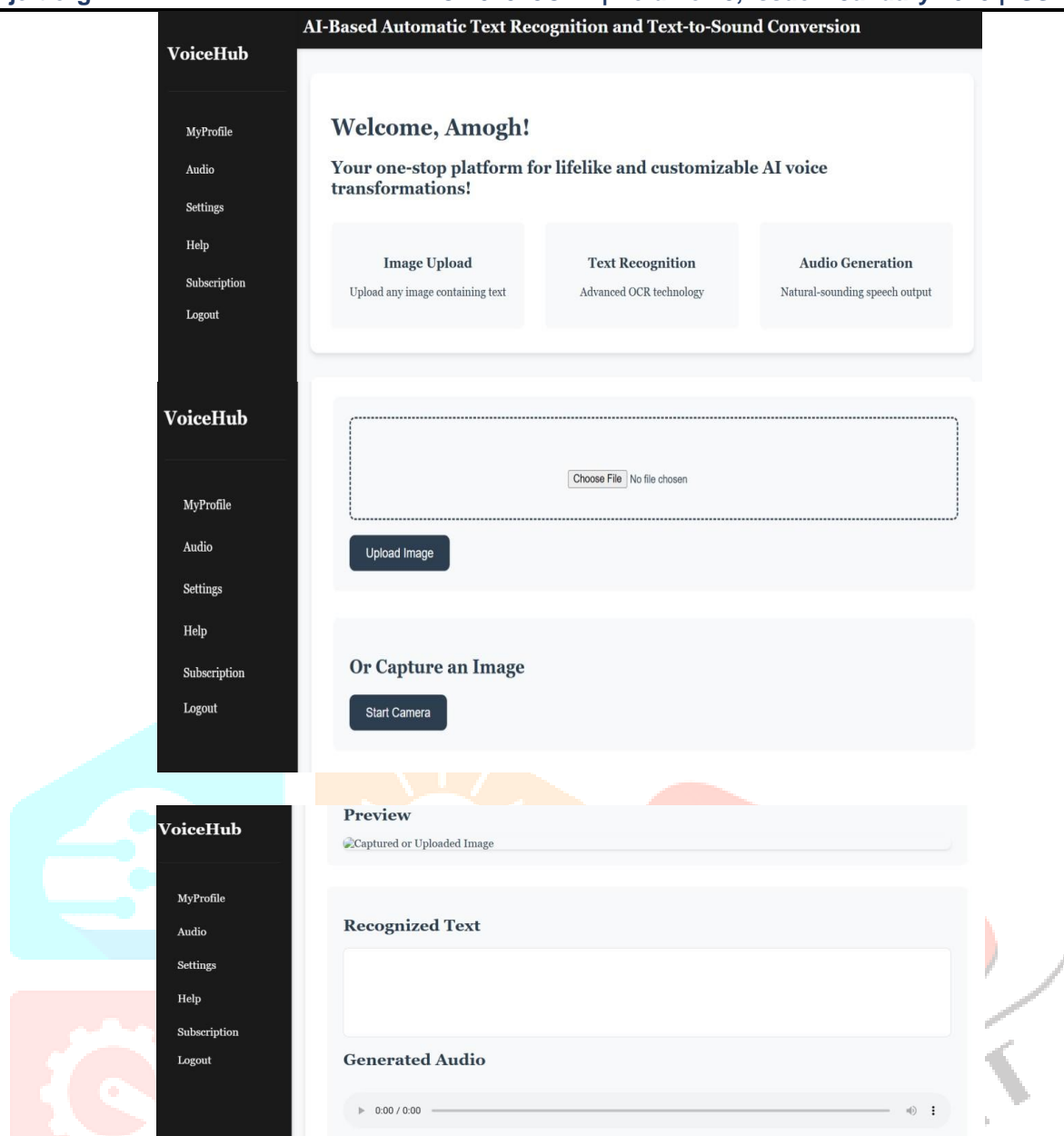


Fig: After logging in, users are redirected to their personalized dashboard, which includes options to upload images, view previous uploads, and access profile settings.

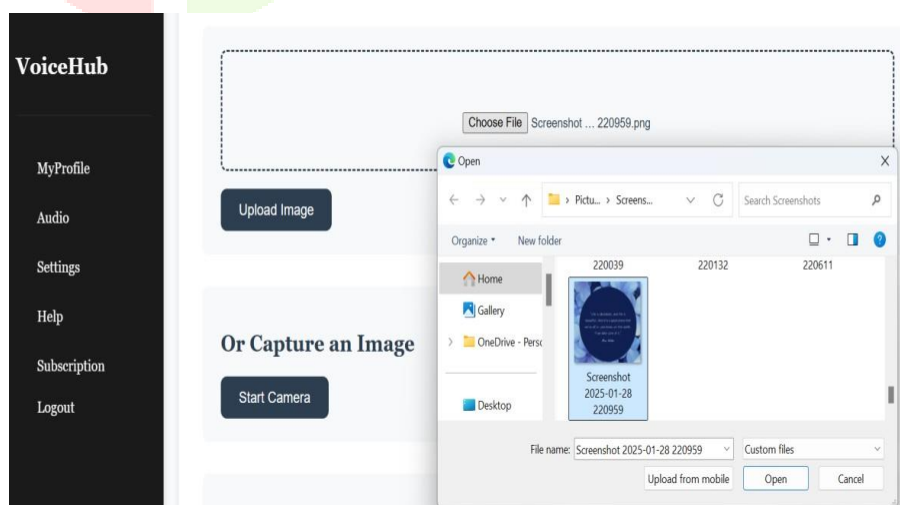


Fig: Users can upload an image or capture one via webcam. The uploaded image is processed using the Google Cloud Vision API to extract text.

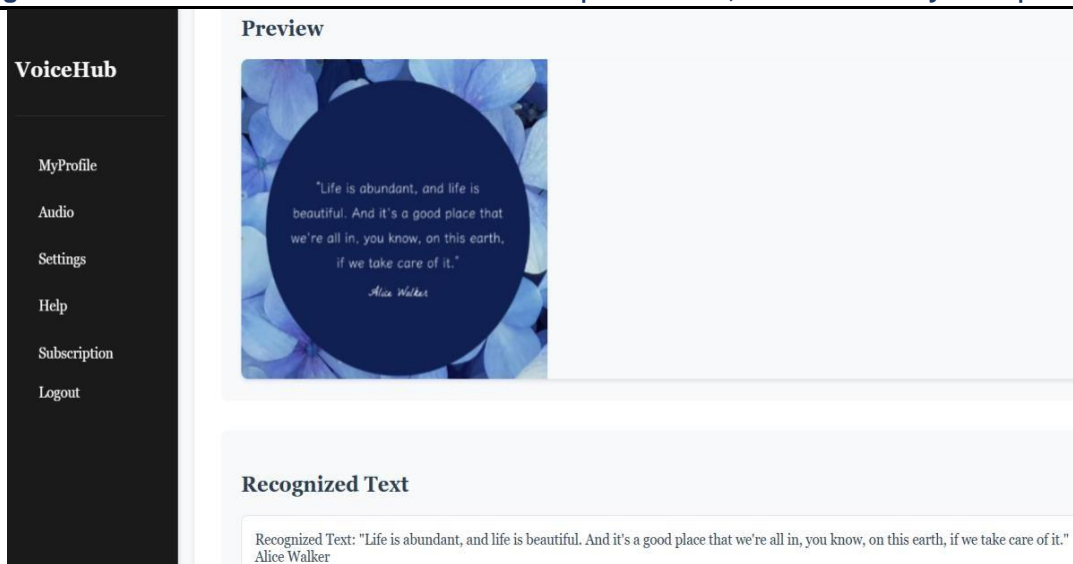


Fig: The system displays the extracted text after processing the uploaded image.

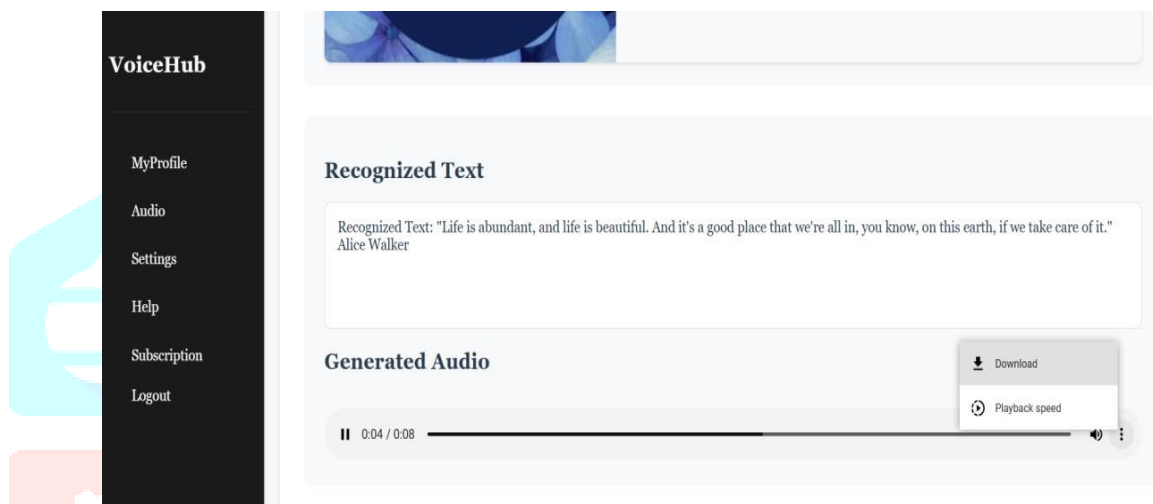


Fig: The recognized text is converted to speech using the Google Cloud Text-to-Speech API. Users can play or download the generated audio file

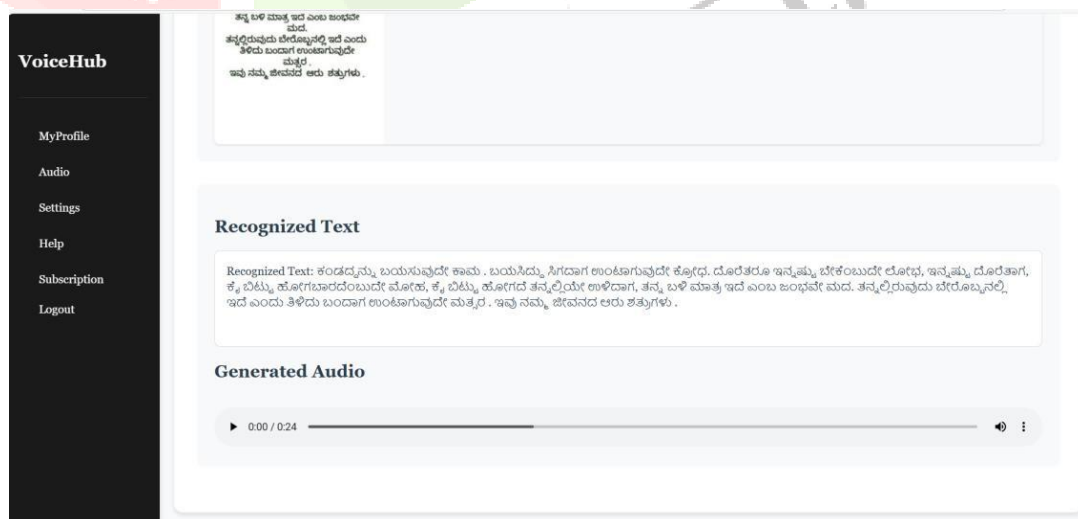


Fig: The system supports text recognition and speech synthesis in multiple languages. An example of Kannada text recognition and conversion is shown below.

Fig: Users can manage their personal details, change their password, and view their activity history from the profile settings section.

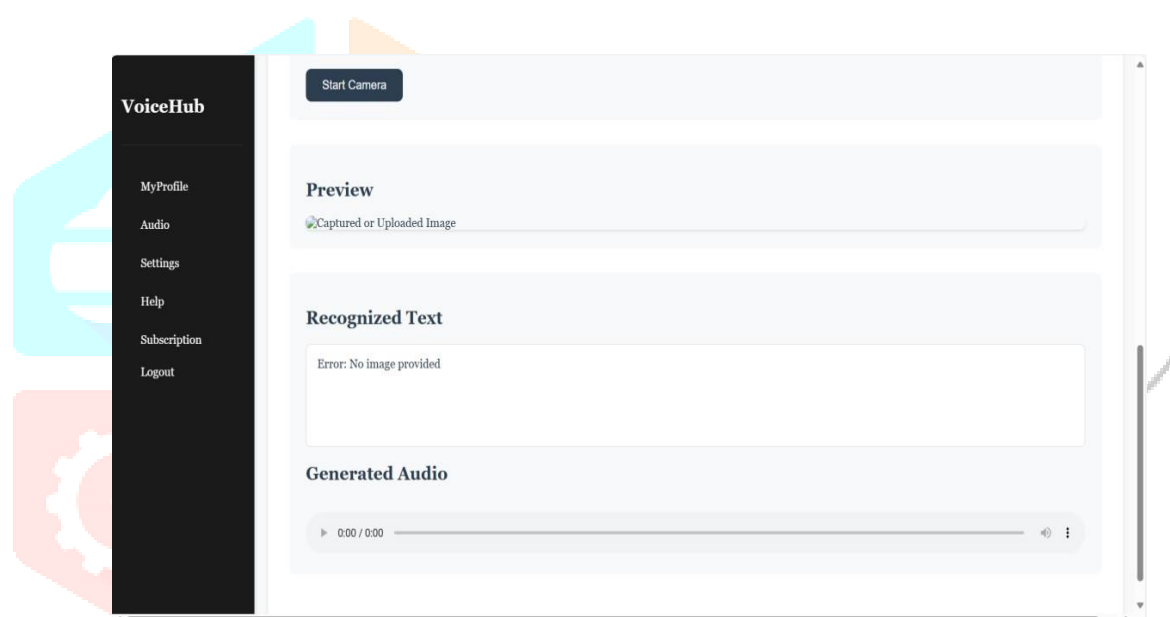


Fig: The system provides real-time error messages for various invalid actions, such as unsupported file uploads, incorrect login details, and duplicate registrations.

## CONCLUSION

The AI-Based Automatic Text Recognition and Text-to-Speech Conversion System successfully integrates cutting-edge technologies to provide a seamless user experience. By leveraging Google Cloud Vision API for Optical Character Recognition (OCR) and Google Cloud Text-to-Speech API for audio generation, the system offers real-time solutions for text extraction and audio playback across multiple languages and formats.

This project demonstrates the effectiveness of AI-driven solutions in addressing real-world challenges, such as text digitization and accessibility. The robust architecture ensures secure user interaction, efficient data processing, and minimal latency, making it suitable for diverse user groups, including those with specific accessibility needs. Through this initiative, we have gained practical experience in implementing AI technologies, handling API integrations, and ensuring user-centric design. The challenges encountered, such as optimizing image preprocessing and handling mixed-language texts, have provided valuable insights into the complexities of deploying intelligent systems in production environments.

Overall, this system lays the foundation for future enhancements, such as support for additional languages, more advanced speech synthesis, and scalability to cater to larger audiences. It highlights the transformative potential of artificial intelligence in simplifying everyday tasks and creating inclusive technological solutions.

## REFERENCES

- [1] **Google Cloud Vision API.** (n.d.). *Google Cloud Documentation*. Retrieved from <https://cloud.google.com/vision>
- [2] **Google Cloud Text-to-Speech API.** (n.d.). *Google Cloud Documentation*. Retrieved from <https://cloud.google.com/text-to-speech>
- [3] **Reddy, S. M., Rao, R. V., & Subrahmanyam, S. V.** (2023). *A Comprehensive Review on Text-to-Speech Conversion Systems*. *Journal of Applied Research on Information Technology*.
- [4] **Patel, J., Kumar, M. D., & Sharma, P. R.** (2022). *Optical Character Recognition (OCR) and Its Application in Real-Time Text-to-Speech Systems*. *International Journal of Computational Vision and Engineering*.
- [5] **Yadav, S., Jha, S. K., & Sharma, A.** (2021). *A Survey on Deep Learning Approaches for Optical Character Recognition*. *International Journal of Machine Learning and Computing*.
- [6] **Zeng, F., Hu, L., & Zhang, Y.** (2021). *Advances in Speech Synthesis: From Concatenative to Neural Networks*. *Journal of Speech Technology*.
- [7] **Pandey, R. S., Tripathi, A. K., & Mishra, N. P.** (2020). *Real-Time Multilingual OCR and Text-to-Speech System for Indian Languages*. *International Conference on Multilingual Computing*.
- [8] **Kumar, A., & Sharma, P.** (2022). **Optical Character Recognition for Indian Scripts Using Machine Learning Techniques**. *Journal of Indian AI Research*.
- [9] **Singh, R., & Gupta, P.** (2021). **Multilingual Text-to-Speech Systems for Indian Languages: A Review**. *Indian Journal of Information Processing and Speech Technology*.
- [10] **Prasad, N. V., & Rao, P.** (2020). **OCR and Text-to-Speech Integration for Regional Indian Languages**. *Proceedings of the National Conference on Emerging AI Technologies in India*.
- [11] **Ramesh, G., & Bhat, S.** (2021). **Deep Learning-Based OCR for Kannada and Other South Indian Languages**. *Journal of South Asian Computational Linguistics*.